

A generative model for protein contact networks

Lorenzo Livi^{*†1}, Enrico Maiorino^{‡2}, Alessandro Giuliani^{§3}, Antonello Rizzi^{¶2}, and Alireza Sadeghian^{||1}

¹Dept. of Computer Science, Ryerson University, 350 Victoria Street, Toronto, ON M5B 2K3, Canada

²Dept. of Information Engineering, Electronics, and Telecommunications, SAPIENZA University of Rome, Via Eudossiana 18, 00184 Rome, Italy

³Dept. of Environment and Health, Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Rome, Italy

March 10, 2015

Abstract

In this paper we present a generative model for protein contact networks. The soundness of the proposed model is investigated by focusing primarily on mesoscopic properties elaborated from the spectra of the graph Laplacian. To complement the analysis, we study also classical topological descriptors, such as statistics of the shortest paths and the important feature of modularity. Our experiments show that the proposed model results in a considerable improvement with respect to two suitably chosen generative mechanisms, mimicking with better approximation real protein contact networks in terms of diffusion properties elaborated from the Laplacian spectra. However, as well as the other considered models, it does not reproduce with sufficient accuracy the shortest paths structure. To compensate this drawback, we designed a second step involving a targeted edge reconfiguration process. The ensemble of reconfigured networks denotes improvements that are statistically significant. As a byproduct of our study, we demonstrate that modularity, a well-known property of proteins, does not entirely explain the actual network architecture characterizing protein contact networks. In fact, we conclude that modularity, intended as a quantification of an underlying community structure, should be considered as an emergent property of the structural organization of proteins. Interestingly, such a property is suitably optimized in protein contact networks together with the feature of path efficiency.

Keywords— Protein contact network; Generative model; Graph Laplacian; Mesoscopic analysis.

1 Introduction

Protein contact networks (PCNs) are minimalistic models of protein 3D structures, which collapse the full-rank information of 3D coordinates of each atom into an adjacency matrix providing the pairwise contacts between residues [14, 21, 36–39, 49, 55, 60, 66]. A contact is scored if the Euclidean distance between alpha carbons of each residue pairs is within two van der Waals radii. PCNs allow for a reliable reconstruction of the global protein structure [62]. Moreover PCNs allow for an efficient description of relevant biological

^{*}llivi@scs.ryerson.ca

[†]Corresponding author

[‡]enrico.maiorino@uniroma1.it

[§]alessandro.giuliani@iss.it

[¶]antonello.rizzi@uniroma1.it

^{||}asadeghi@ryerson.ca

properties of proteins such as allosteric effect and identification of active sites [20]. The efficiency of PCNs in retaining the essential features of protein structures makes the development of a PCN generative model of utmost importance for shedding light on both folding process and the structural bases of the unique functional properties of protein molecules [4, 24, 50, 64]. Although many generative models have been developed in the still young network science discipline [7, 47], fewer and less established examples are available in the literature when focusing on formal representations of protein molecules [2, 5, 8, 23, 43, 53, 54, 58]. It is possible to cite approaches for modeling proteins based on knot theory [43] and topology [53, 54, 58]. Other approaches focus on approximating particular protein structures, such as the local structure [8] and specific fold families [2]. To the best of our knowledge, only Refs. [5, 23] studied the problem of generating PCNs for evaluating detailed graph-theoretical characteristics.

The quest for a reliable and, most importantly, justifiable generative model for PCNs implies as a first step the identification of a target function. This allows for a unambiguous evaluation of the proposed model in terms of “superposition” of the simulated contact networks with the real PCNs. Inspired by the seminal works by Leitner [34], we considered here as target function to approximate of the peculiar heat trace decay of PCNs [38]. Such a property is elaborated from the heat kernel [11, 31, 65], the graph-theoretical analogue of the well-known first-order differential equation describing diffusion of heat in a physical medium. The heat trace (HT) is an invariant property – in the graph-theoretical sense – elaborated from the spectrum of normalized graph Laplacian [1, 41, 42, 44]. Graph Laplacians are objective of focused studies in many scientific fields, as in fact it is possible to extrapolate many structural and dynamical properties from such a matrix representation of the system [18, 33, 52, 67]. The interest in the study of graph Laplacians motivated also related researches focusing on the so-called spectral reconstruction, i.e., on calculating a graph given a specific spectrum to be considered as target [12, 28]. Although potentially interesting, co-spectrality of graphs is a very hard problem and it is still a not very well-developed field in graph theory [61], limiting hence its practical exploitation.

Inside a protein molecule, energy readily flows between distant regions connected by inter-module contacts (fast lane) and only slowly within modules (slow lane). In other words, proteins present either a strong modularity, causing a low thermal dissipation (heat is kept into modules by the richness of dead ends pathways slowing down the spreading of energy), and a suitable number of long-range shortcuts, allowing for a rapid and yet efficient communication between distant sites responsible for allosteric effect [34]. This dual behavior is at the basis of two crucial properties for protein physiology: 1) keeping a stable micro-environment for the catalyzed reaction (slow decay) and 2) allowing for an efficient spreading of information throughout the molecule to get rid of environmental changes (binding of an allosteric effector, pH changes, etc). There is a clear trade-off between the two above goals: increasing modularity is good for the first goal but is detrimental for the latter; on the other hand characteristic length minimization is an efficient strategy for the latter but is detrimental for the former.

The contribution of this paper consists in a two-step generative model for PCNs; the first stage of our method takes inspiration from the work of Bartoli et al. [5]. The dataset considered in our study consists of four ensembles (classes) of networks: i) actual PCNs elaborated from the *E. coli* proteome [36, 37], ii) synthetic networks generated according to the recipe of Bartoli et al. [5], iii) synthetic modular networks generated with the method proposed by Sah et al. [57], and finally iv) those generated with our method. We evaluate the soundness of the proposed approach by focusing on mesoscopic analyses. Notably we first study characteristics elaborated from the normalized Laplacian spectra. To complement this spectral analysis, we also study topological descriptors, including statistics of the shortest paths and quantification of the network modularity. Results show that the ensemble of networks generated with our method ends up in a statistically significant improvement of similarity with real PCNs, as for both spectral and topological properties. However, a principal component analysis (PCA) of the considered topological descriptors revealed a gap with actual PCNs, specifically related to the shortest paths. The second step of the proposed method is hence designed to compensate this drawback. A new version of the ensemble of networks is then obtained by rewiring edges with high edge-betweenness [16]. As a result, we show that we are able to achieve a further statistically significant improvement of the ensemble characteristics, without altering the global spectral properties of the first ensemble that we generated.

As a byproduct of our study, we demonstrate that modularity, a well-known feature found in proteins as well as in many other biological networks, is not sufficient to explain the underlying network architecture of PCNs. This result is of particular interest, since it stresses the peculiar architecture of proteins that suitably merges conflicting features such as path efficiency and modularity. The topological properties of the obtained graphs, exploited by PCA, allowed us to offer a clear structural counterpart to HT features. These structural counterparts are by no means confined to protein science, given that well-known relations between small-world, fractal, and modular architectures with path efficiency [25, 56] are present also in brain networks [9] and are likely to constitute relevant features to be optimized in other natural and/or artificial networked systems.

The remainder of this paper is structured as follows. In Section 2 we describe the data that we considered in our study. In Section 3 we show and discuss the obtained results. Section 4 offers the conclusions and pointers to future research works. The proposed generative algorithm is fully described in the Methods section, notably in Section A.1; the considered graph characterizations are instead introduced in Section A.2.

2 Dataset

In our study we consider four ensembles of networks, three of which are created according to specific generative models. Each ensemble contains 100 networks of varying size (from 300 to 1000 vertices). Each graph with n vertices and m edges is present in the four different ensembles preserving n and m , but varying the resulting topology according to a generative mechanism. This is performed to focus the analysis on ensemble features proper of the structural organization, without considering effects due to the size of the graphs.

The first ensemble of graphs contains PCNs, directly obtained from the 3D native structures resolved for the E. coli proteome [36, 37]. Each vertex is defined as the alpha carbon of a residue; edges are added among two residues if their Euclidean distance is within the $[4, 8]$ Å range. This choice is justified by observing the densities of native contact length in Fig. 1, respectively when considering all contacts below 8 Å and the herein adopted filtered version. It is possible to note a peak right before 4 Å, which corresponds to trivial contacts due to closeness of the residues on the backbone [21]; such contacts do not alter the typical network architecture of PCNs and therefore they are not considered in our PCN graph representation.

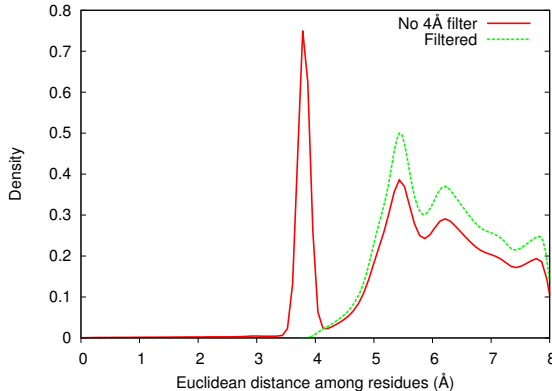


Figure 1: Density of Euclidean distances among native contacts in PCNs.

The second ensemble of networks is elaborated by using the recipe of Bartoli et al. [5], which considers a set of edges added deterministically due to the backbone and additional edges inserted according to a probability that scales linearly with the sequence distance of residues. For the third ensemble we use networks generated according to the recently-proposed scheme by Sah et al. [57]. Such a generation mechanism produces modular networks with controlled (i.e., user-defined) modularity and degree values. Such a generation mechanism is inspired by the fact that modular structures seem to be ubiquitous in biological networks; modularity is

considered to be at the basis of resilience and adaptability features of biological networks. It is well-known that PCNs are modular, i.e., they possess a well-defined community structure [38]. Accordingly, the third ensemble of Sah et al. networks is generated by copying modularity and degree from the considered PCNs. The contemporary presence of networks in which modularity (if any) comes up as an emergent property and Sah et al. networks, in which modularity is a built-in property, will help us to check if the peculiar PCN spectral and topological properties are a direct consequence of their pronounced modularity or not. Finally, the fourth ensemble of networks is constituted by 100 graphs generated according to the herein proposed mechanism, referred to as LMGRS networks, whose first step consists in a variant of the Bartoli et al. model. In this adaptation, the linear scaling of the probability of the edges with the distance in sequence is replaced by the empirical frequencies measured in the PCN ensemble. As it is discussed later, the LMGRS are successively reconfigured in an iterative fashion to obtain a new ensemble of networks, shortened in the following as LMGRS-REC. The proposed LMGRS generation and LMGRS-REC reconfiguration methods are fully described in Sec. A.1.

3 Results

The four ensembles are first evaluated in terms of mesoscopic properties elaborated from the spectra of the normalized graph Laplacians (see Sec. A.2 for technical details). Fig. 2 shows, respectively, the characteristic HT decay and the ensemble spectral densities. The HT decay analysis (2(a)) offers insights on an ensemble in terms of characteristic diffusion time. The analysis takes into account the varying-size character of the considered networks. From the plot it is possible to note that LMGRS introduces a considerable improvement with respect to Bartoli et al. and Sah et al. ensembles: the former decays around $t \simeq 350$ while the latter decay much faster at $t \simeq 100$. However, the PCN trend is not accurately approximated yet. This improvement of several orders of magnitude in the characteristic HT decay time can be explained by focusing on the spectral densities shown in Fig. 2(b). LMGRS ensemble density possesses clear similarities with the one of Bartoli et al., being the two based on the same algorithmic template. Nonetheless, by highlighting the lower band of the spectra, it is possible to note some important differences for a specific region (in-between 0 and 0.2) containing eigenvalues related to the modular structure of networks. LMGRS ensemble offers a better approximation for those small eigenvalues, which explains the significant improvement observed for the HT decay.

Now we move to the analysis of the ensembles by considering the representation of each graph as a numeric vector containing suitable features that characterize different aspects of the network topology (see Sec. A.2). To offer a synthetic visualization, in Fig. 3 we show the pairwise relations among the first three principal components (PC) elaborated via the principal components analysis (PCA) of such a vector-based representation of the considered networks; data is standardized prior to PCA processing. The first three PCs explain $\simeq 92\%$ of the entire data variance (PC1 $\simeq 39\%$, PC2 $\simeq 30\%$, and PC3 $\simeq 23\%$) and therefore they are considered as the signal part of information; the component loadings (Pearson correlation coefficients between original descriptors and components) are reported in Tab. 1. Loadings on the PC offer a particularly interpretable scenario, where PC1 is mostly explained by the path distribution (ACC and ASP) and the local clustering (ACL). As expected ACC negatively scales with both ASP and ACL, so pointing to the fact that ACL decreases the efficiency of signal transmission across the network (positive correlation with ASP). Thus, high values of PC1 corresponds to architectures with elevated characteristic length (slow information transmission), while low values of PC1 point to graphs with elevated closeness centrality (ACC) and thus relatively efficient signal transmission. PC2 is mainly correlated with MOD, A and H, with A going in the opposite direction with respect to the other two descriptors. This corresponds to the fact the regularity of a graph decreases as the modularity increases; it is also well-known that modularity affects random walks behavior, explaining the positive correlation with H. PC3 is entirely described by the spectra of the adjacency and Laplacian matrices (respectively indicated by EN and LEN).

The PCs are linearly independent by construction, so the above results clearly indicate that the dataset can be described by three autonomous topological features: 1) path length and local clustering (PC1); mesoscopic modularity (PC2); and 3) spectral properties (PC3). The particular mixing of these independent

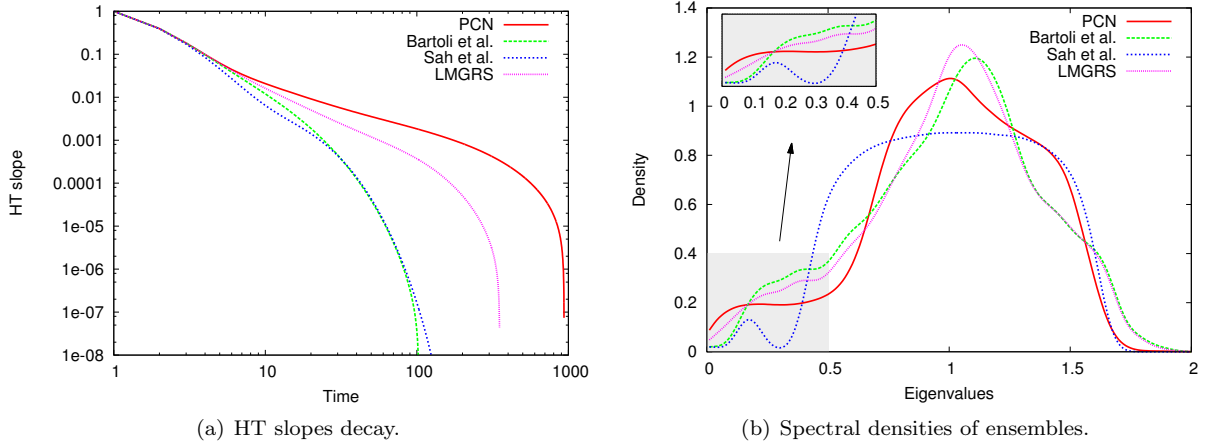


Figure 2: Ensemble HT slopes decay (13) for the considered graphs 2(a) and related Laplacian spectral densities 2(b). The proposed LMGRS model clearly denotes more similar characteristics with respect to PCNs in terms of HT decay. Analogously, the LMGRS model induces a density of eigenvalues more similar to PCNs in the lower bands (see detailed plot), suggesting that the community structure is more suitably approximated. The Sah et al. model instead does not mimic the spectral density, even if the modularity level and degrees have been copied from the original PCNs.

features varies across the different ensembles. Let us now focus on the PCA subspace spanned by PC1–PC2 (Fig. 3(a)). It is possible to note that LMGRS ensemble introduces an improvement in PC1, which as explained before, encodes contributions in terms of path distribution and local clustering. A very interesting scenario can be observed when considering the projection given by PC1–PC3 (Fig. 3(b)). In fact, when PC2 is not considered Sah et al. and LMGRS become very similar to each other, and entirely different from the ensemble of Bartoli et al. To summarize, it is worth pointing out that the average Euclidean distance among the LMGRS and PCNs networks represented in the three-dimensional PCA space is significantly inferior ($p < 0.0001$) with respect to the distances among Bartoli et al. and PCNs (3.13 vs 4.69 with standard deviations 0.75 and 0.68, respectively).

Table 1: Principal component loadings.

	PC1	PC2	PC3
MOD	0.26148	0.78872	0.13409
ACC	-0.97835	-0.15695	-0.11588
ASP	0.92838	0.26494	0.09270
ACL	0.89098	-0.22533	-0.12227
EN	0.01862	0.27707	0.95810
LEN	0.04354	-0.27981	0.94213
H	0.05171	0.99519	-0.04393
A	0.09073	-0.84835	0.06463

From PCA space snapshots we deduce that the LMGRS ensemble is a considerable improvement with respect to the others, while there is still a gap to be filled with respect to PCN. Notably, LMGRS networks present a too small characteristic path length – the small-world signature is too strong. This fact explains the differences observed for what concerns path distribution and modularity, since in fact the path efficiency and modular properties are two conflicting features in networks, which usually survive in the same network

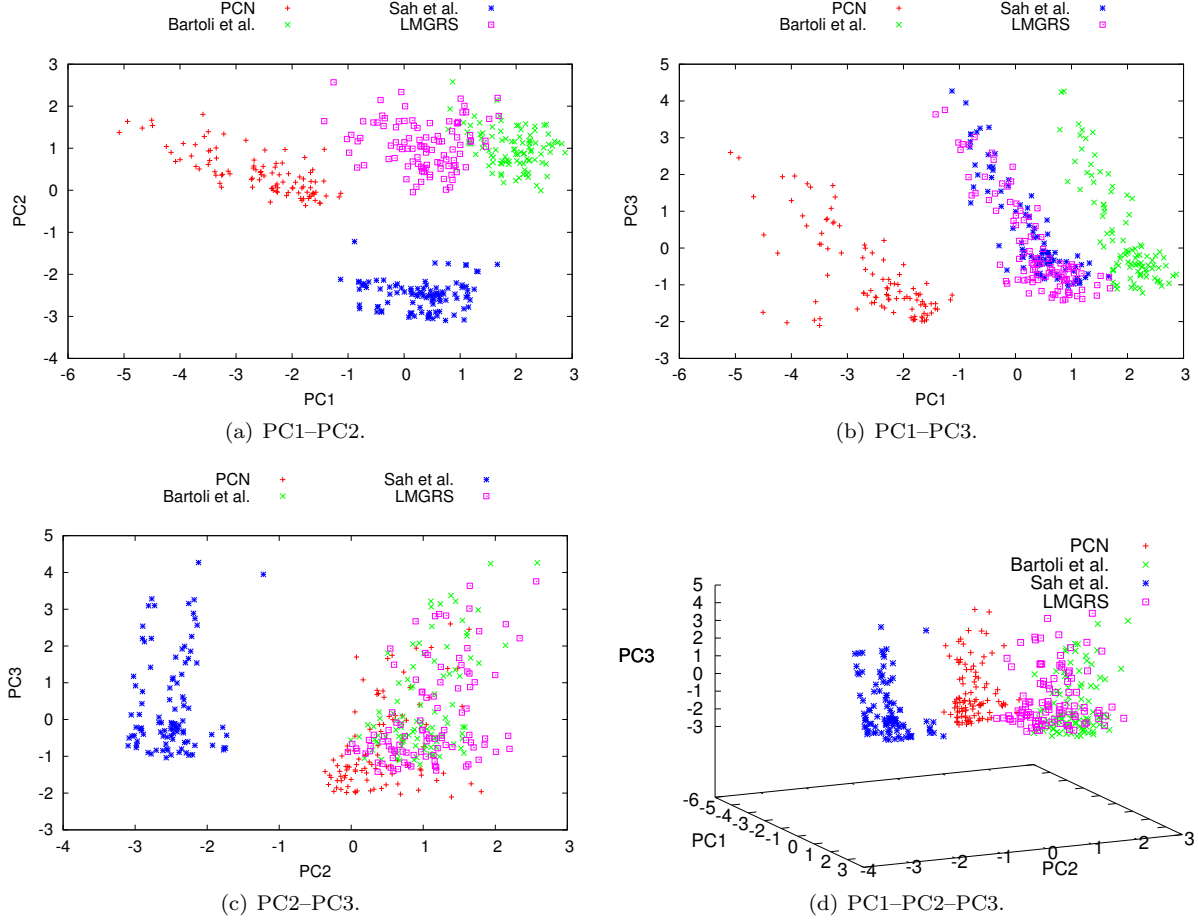


Figure 3: PCA of the topological descriptors calculated on the four ensembles of protein graphs. LMGRS model is an improvement with respect to Bartoli et al. [5] also considering classical TD.

in terms of a trade-off (modular organization is progressively lost as the network becomes more and more efficient in terms of shortest paths). To this end, as mentioned earlier, we consider another ensemble derived by post-processing LMGRS networks with the edge reconfiguration process described in Sec. A.1, denoted as LMGRS-REC. The reconfiguration process is targeted to rewire edges with high edge-betweenness, since those edges have a direct impact on path efficiency, and accordingly also on the modular organization. Each graph of the LMGRS ensemble is reconfigured according to the following process (see Sec. A.1 for details). At each iteration, the edge with maximum edge-betweenness is removed and it is re-attached to two randomly chosen vertices at a backbone distance given by the empirical distribution shown in Fig. 9. This process is repeated until a suitable convergence criteria is met. In our case, we considered a number of iterations (50) that resulted in a statistically significant improvement of the ensemble features that we observe. In Fig. 4(a) we show the detailed changes of the TD for the LMGRS and LMGRS-REC with respect to the PCNs. The figure reports, for each descriptor, the average absolute difference calculated for each graph of the respective ensembles with respect to the PCN graphs; standard deviations are reported as vertical bars. Results show that the reconfiguration algorithm performs well as for statistical significance of differences with PCN, assessed via t-test with the usual 5% threshold. In particular, as desired reconfigured networks denote more similar ASP ($p < 0.0025$) and ACC ($p < 0.0001$). As expected, such improvements for the shortest paths have a direct influence on the global modularity. In fact, MOD is significantly improved ($p < 0.0018$). This is a direct consequence of the fact that path efficiency and modularity are features to be

considered in a trade-off. ACL similarity improves as well ($p < 0.0059$), denoting a better approximation of the local cluster structure of PCNs. It is important to note that differences for EN ($p < 0.0621$) and LEN ($p < 0.1134$) are not statistically significant (especially those for LEN). This fact tells us that spectral properties of the reconfigured networks are not significantly altered. Fig. 5(b) offers a visual confirmation of this fact. In fact, the spectral densities for LMGRS and LMGRS-REC reported in the figure are almost identical. However, it is worth discussing the HT slopes shown in Fig. 5(a). From the figure, it is possible to notice a slight divergence among LMGRS and LMGRS-REC for large-time instants. This is due to the difference in magnitude of the first non-zero eigenvalue of the normalized Laplacian, which particularly influences the asymptotic HT behavior. Such a difference is a byproduct of the designed edge reconfiguration algorithm, which focuses on rewiring edges with high edge-betweenness: those are most likely connections among different densely connected communities. In graph-theoretical terms, LMGRS-REC networks are characterized by a lower conductance with respect to the LMGRS ensemble. In fact, the conductance is directly related to first non-zero eigenvalue of the Laplacian (details not shown here). Finally, differences for both H ($p < 0.0030$) and A ($p < 0.0069$) are significant as well.

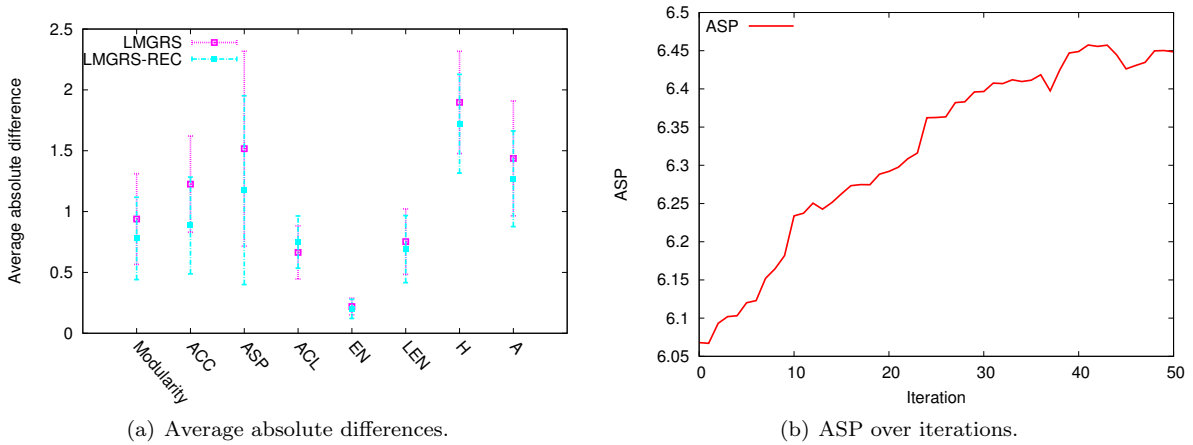


Figure 4: Average differences for each TD with respect to PCN (4(a)) and their standard deviations. We are able to modify, among the other factors, the small-world character of the generated networks without significantly affecting the spectra of the adjacency (EN) and Laplacian (LEN) matrices. Statistical significance of differences is assessed via t-test. Fig. 4(b) shows the ASP of a sample graph during the reconfiguration process.

To conclude this section, we discuss the results shown in Fig. 6. The figure shows the PCA performed by considering the (standardized) HC coefficients (5) of the first 100 time instants. HC coefficients include, in addition to the spectra, also the information of the eigenvectors of the normalized Laplacian, which encode the arrangement of the vertices in a given vector space. In the literature (see Ref. [44] and references therein) this is called localization of the eigenvectors and it is usually exploited for clustering purposes. While variance is almost entirely explained by PC1 ($\simeq 99.8\%$), for our purposes we consider the first three PCs. From PC1–PC2 it is possible to note that PCN, LMGRS, LMGRS-REC, and Sah et al. have a very well-defined and compact configuration in the PCA subspace. Since Sah et al. possesses a well-defined community structure, we deduce that when considering the information of the eigenstructure of the normalized Laplacians, all four ensembles (i.e., PCN, LMGRS, LMGRS-REC, and Sah et al.) possess striking similarities. However, Bartoli et al. ensemble denotes a very disorganized configuration regardless of the considered PCA subspace (i.e., PC1–PC2 or PC2–PC3), which could be intended as a symptom of a weaker modular organization of the vertices.

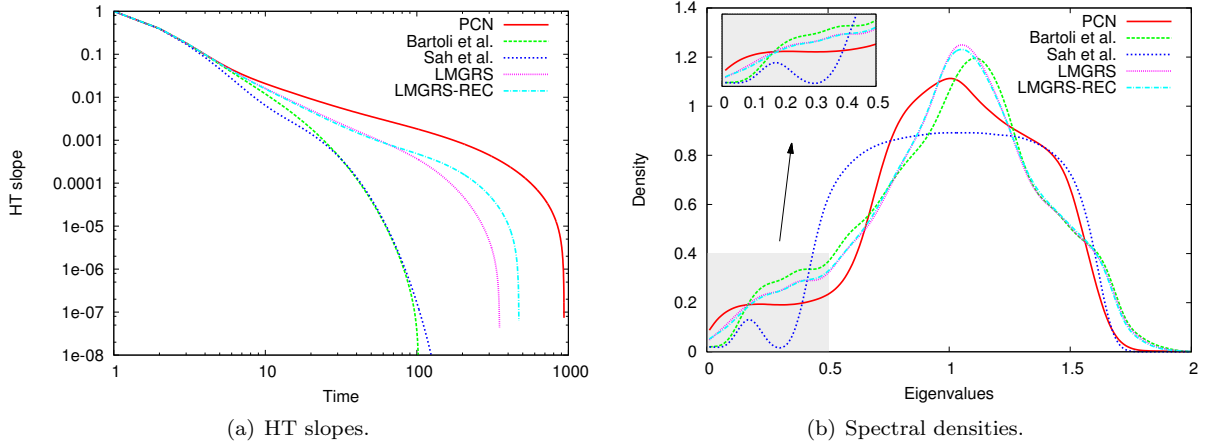


Figure 5: Same as Fig 2 but including also reconfigured LMGRS.

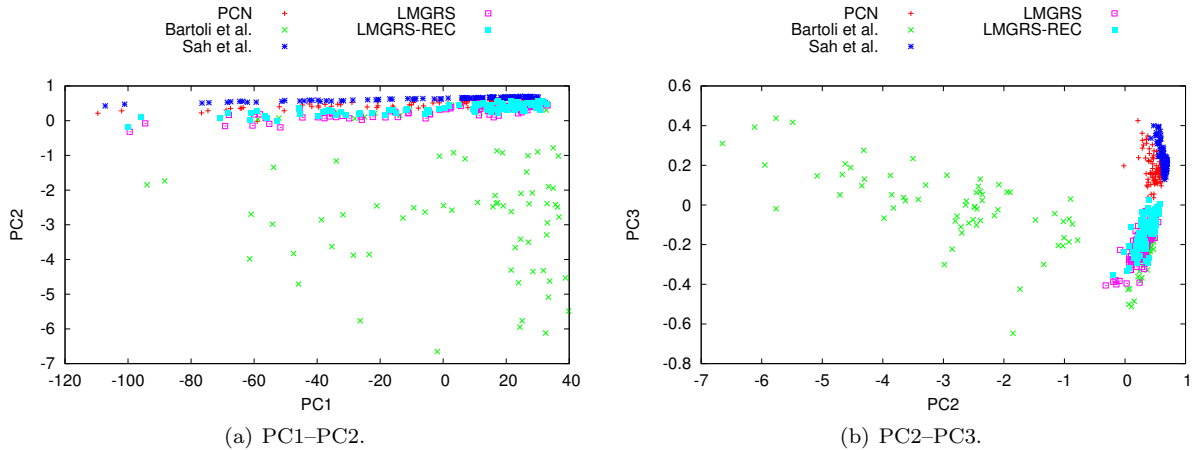


Figure 6: PCA of the first 100 time instants of the HC coefficients elaborated from the ensembles.

4 Conclusions and future directions

In this paper, we have proposed a two-step generative model for protein contact networks. For the first step, we partially took inspiration from the work of Bartoli et al. [5], whose idea was to generate contact graphs by first adding backbone contacts deterministically (considering adjacent residues along the sequence). Successively, a number of additional contacts were added with a probability inversely proportional to the residue distance along the sequence. Here we modified this part by considering the actual empirical probability distribution of contacts with respect to the sequence distance, derived from an ensemble of *E. coli* proteins. We analyzed our generative method by considering three additional ensembles composed of 100 varying-size protein contact networks. We focused on a mesoscopic analysis, that is, we primarily investigated the soundness of the models by considering information derived from the eigendecomposition of the normalized Laplacian. Results showed that the proposed method approximates with better precision the behavior of actual protein contact networks in terms of characteristic diffusion time. We considered also several common topological descriptors. This last analysis pointed out that our method, as well as the others, does not approximate sufficiently well the path distribution. To this end, we designed an edge reconfiguration algorithm to be used as the second step of the proposed generative method. We then generated an additional ensemble of

reconfigured networks, which showed statistically significant improvements with respect to the initial one.

We considered an ensemble composed of graphs synthesized according to a recently-proposed mechanism [57], designed to construct a network with specified modularity and degree profiles. Notably, we reproduced the modularity and degree values from the actual protein contact networks herein considered. Results demonstrate that modularity, when hardcoded into the networks, does not explain the actual architecture of proteins. In fact, we concluded that modularity should be considered as an emergent property of such networks, which is suitably optimized in a trade-off with the conflicting feature of path efficiency. In our model, modularity emerged from the peculiar PCNs mesoscopic wiring obtained from their empirical contact distribution at increasing distance length: a simple linear decrease in contact frequency at increasing sequence distance does not allow to reach the typical modularity of real proteins. The fine tuning of long-range contacts allows for directly intervening on both modularity and path efficiency balance, so confirming the crucial importance of long-range contacts in folding process [10, 63].

A sound generative mechanism for protein contact networks is of utmost importance in current researches in protein science. In future works, in addition to the improvement of the herein proposed method, we also plan to tackle the problem in a data-driven generative learning scenario, for instance using generative (deep) neural networks [32]. The possibility to learn in a data-driven fashion an effective and sound model for protein contact networks would allow to easily generalize other instances of such networks. This perspective could be interesting also for protein engineering purposes [15]. The theoretical study of networks promises to pave the way for the discovery of universal principles at the basis of biological organization as well as instructing the generation of technological devices.

A Methods

A.1 The proposed generative method for synthesizing protein contact networks

In this section we describe the proposed generative method for PCN. Instead of merely presenting the mechanism itself, we first discuss an important fact related to the distance on sequence of amino acid residues and their relations on the native 3D structure of proteins. This initial discussion, in our opinion, is relevant for the purpose of designing and, most importantly justifying, a generative mechanism for PCN.

Native contacts in folded proteins are in one way or another constrained by the covalent bonds due to the backbone. Therefore, a first interesting question that one would ask when designing a generative mechanism is “what is the effect of the backbone on the degree distribution of a PCN”. To provide an answer to such a question, we first define the notion of short range (SR) and long range (LR) contacts, that is, native contacts whose residues are, respectively, close and distant on the sequence (backbone). We chose 12 residues as threshold for SR contacts [51]. Fig. 7 shows the two separate degree distributions elaborated from the considered ensemble of varying-size PCNs. SR contacts denote a clearly different distribution with respect to those that are LR; the latter is vaguely compatible with a power-law.

Considering this fact and that PCNs do possess a modular architecture, one would be tempted to postulate a striking rule such as “SR contacts are intra-module while LR are inter-module links”. If this rule would be correct, it would be possible to design a generative mechanism accordingly, e.g., by connecting intra-module and inter-module links according to their specific (empirical) distributions; although the number of modules should be defined a priori. Nevertheless, such a possibility seems to be weakly supported by the following test. In Fig. 8 we show a graphical representations of two PCNs, denoted as “JW0058” and “JW0179”. Those two networks contain roughly the same number of amino acid residues (around a thousand); JW0058 is made of two chains while JW0179 is derived from a single-chain polymer. To verify the above stated hypothesis, we need to consider a suitable criterion to generate a partition (i.e., to group the vertices into modules). In our case, we have chosen to use the partition having maximum modularity as computed with the algorithm presented in Ref. [6]. Results in Fig. 8 demonstrate that intra-module links (solid lines) are SR (drawn in red), in both cases, only around 55% of the times. This fact – that has been verified for a larger number of PCN – suggests to reconsider the possibility to follow such a SR/LR contacts characterization with respect to intra/inter module links. In addition, we found in our data that there is no trivial relation among

the distance on sequence and the Euclidean distance among residues in the 3D space ($r \simeq 0.162$, details not shown here). These facts find confirmation by considering the enormous research effort in predicting native contacts starting from the sequence [3, 17, 22, 27, 29, 30, 40, 45, 46, 59].

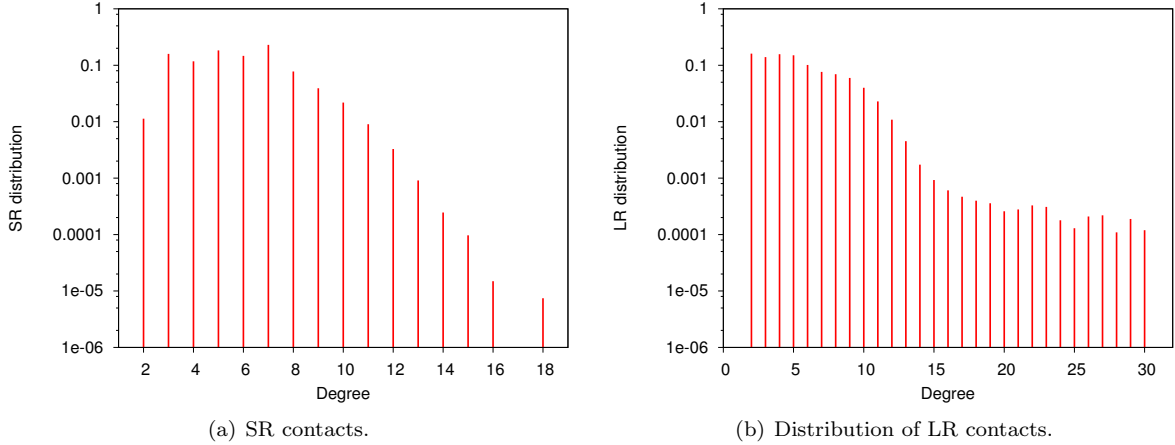


Figure 7: Degree distribution of SR (7(a)) and LR (7(b)) contacts. Both distributions are provided in lin-log plots to improve visualization. SR contacts are determined by considering a distance on the sequence lower than or equal to 12 residues.

Let us describe the proposed generative mechanism. Algorithm 1 conveys the pseudo-code of the procedure. The algorithm takes inspiration from the mechanism introduced by Bartoli et al. [5]. Firstly, edges are deterministically added among any two residues at distance two on the sequence. This provides the definition for backbone contacts. The main difference with Bartoli et al. [5] is that, to add all remaining non-backbone contacts, we use the empirical distribution shown in Fig. 9 instead of a linear function of the sequence distance. As shown in the results, this straightforward modification resulted in a considerable improvement under many aspects.

Algorithm 1 Pseudo-code of the proposed generative algorithm.

Require: Number of vertices, n , and edges, m

Ensure: A graph $G = (\mathcal{V}, \mathcal{E})$ with $n = |\mathcal{V}|$ and $m = |\mathcal{E}|$

- 1: Add n vertices in \mathcal{V} with unique, progressive, numerical identifiers
 - 2: Add backbone contacts in \mathcal{E} : connect all vertices v_i and v_j for which $|i - j| = 2$
 - 3: Loop to add all remaining non-backbone contacts \curvearrowright
 - 4: **while** $|\mathcal{E}| < m$ **do**
 - 5: Select two non-connected vertices v_i and v_j with probability $p(|i - j|)$ given by their distance $|i - j|$ according to the empirical distribution in Fig. 9
 - 6: Add the undirected edge $e = (v_i, v_j)$ in \mathcal{E}
 - 7: **end while**
 - 8: **return** $G = (\mathcal{V}, \mathcal{E})$
-

We now introduce the second step of the proposed generative mechanism, which implements the edge reconfiguration. Given a graph $G = (\mathcal{V}, \mathcal{E})$, the herein introduced reconfiguration step is primarily meant to lower the small-world signature in G . This is done by iteratively rewiring edges in \mathcal{E} according to their edge-betweenness value. The pseudo-code of the edge reconfiguration algorithm is delivered by Algorithm 2. The reconfigured graph \hat{G} is obtained at the end of the reconfiguration loop. Please note that we insure connectedness for all \hat{G} . The loss of small-world signature in \hat{G} is primarily verified with the ASP increase (see Fig. 4(b) for an example), which is a consequence of the targeted rewiring of edges with maximum edge-betweenness.

Algorithm 2 Pseudo-code of the proposed edge reconfiguration algorithm.

Require: A graph $G = (\mathcal{V}, \mathcal{E})$ with $n = |\mathcal{V}|$ and $m = |\mathcal{E}|$

Ensure: A modified graph $\hat{G} = (\mathcal{V}, \mathcal{E})$ with $n = |\mathcal{V}|$ and $m = |\mathcal{E}|$

1: **loop**

2: Calculate the edge-betweenness measure for all edges in \mathcal{E}

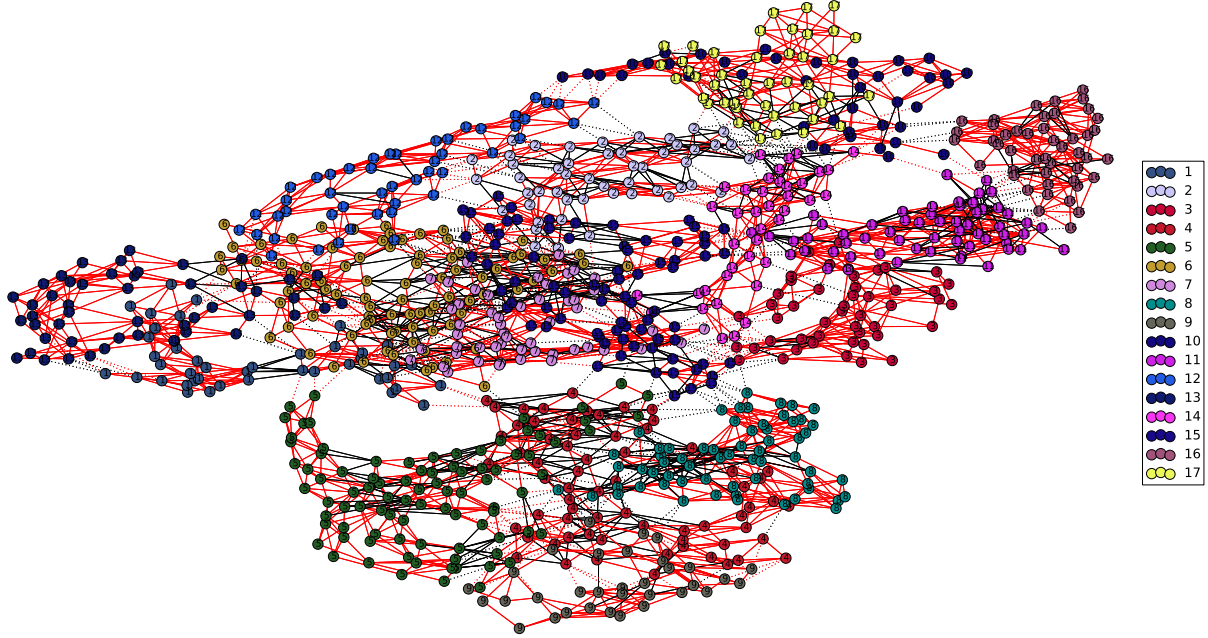
3: Let e_{\max} be the edge with maximum edge-betweenness. Remove e_{\max} from \mathcal{E}

4: Select two non-connected vertices v_i and v_j with probability $p(|i - j|)$ given by their distance $|i - j|$ according to the empirical distribution in Fig. 9

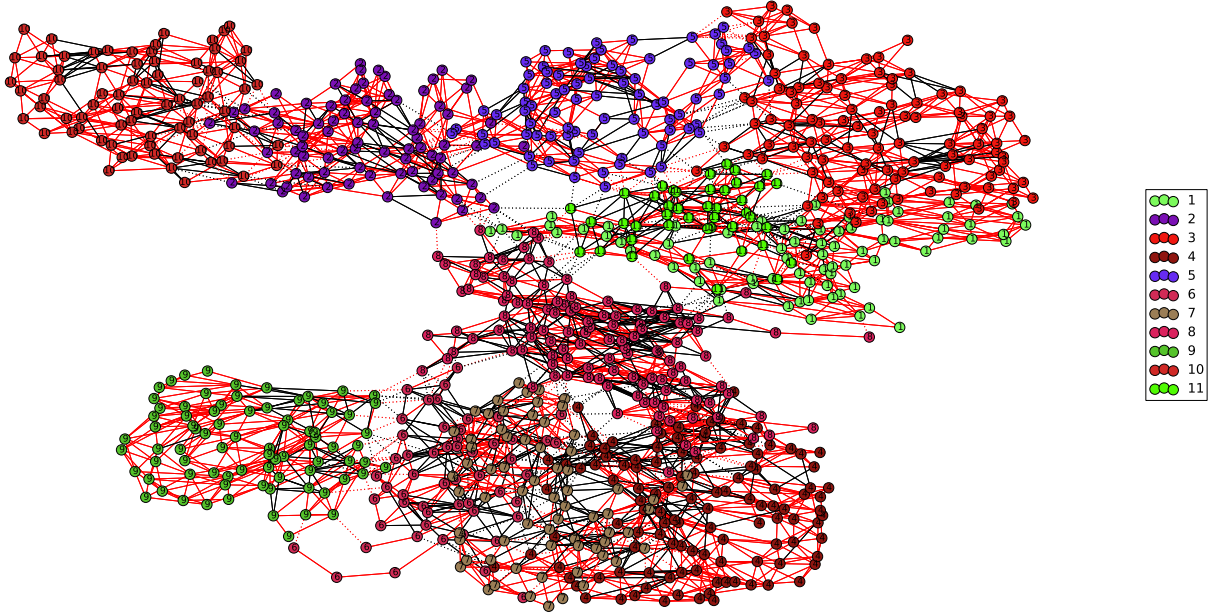
5: Add the undirected edge $e = (v_i, v_j)$ in \mathcal{E}

6: **end loop** when stop criterion is met

7: **return** $\hat{G} = (\mathcal{V}, \mathcal{E})$



(a) JW0058 modules/links organization.



(b) JW0179 modules/links organization.

Figure 8: Classification of contacts by considering the SR/LR typology and the intra/inter module arrangement. The partition is derived by using the maximum modularity criterion. Vertex assignment to modules is represented using different colors; numerical module identifies are drawn in the legend and in the corresponding vertex labels. Solid lines denote intra-module links while dashed lines inter-module links. Black links denote LR contacts, while red links are SR. Please note that the length of the links in the figures respects the actual Euclidean distances of contacts. The assumption that LR contacts are mostly inter-module links (and accordingly, SR contacts are mostly intra-module links) seems to be disproved by those examples.

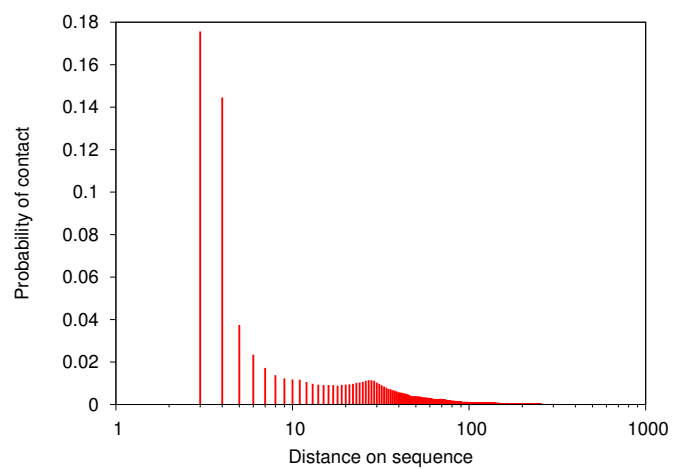


Figure 9: Empirical distribution of contacts considering the distance on sequence (backbone distance). The distribution is elaborated from the entire ensemble of PCN without considering contacts at distance one and two on the sequence; backbone contacts are added deterministically.

A.2 Graph characterization and heat kernel

In this appendix we provide essential technical details regarding the graph characterizations used in this paper to study the four ensembles of variable-size graphs representing proteins.

The first characterization employs classical topological descriptors, which include statistics of the degrees/shortest paths and also some elaborations of the graph spectra. We consider the modularity (MOD) [6, 48] for quantifying the presence of a global community structure – please note that we consider the value associated to the partition with maximum modularity. Then we consider the average closeness centrality (ACC), average shortest path (ASP), and the average clustering coefficient (ACL) [13]; the energy (EN) and Laplacian energy (LEN) of the corresponding spectra [26]; the ambiguity (A) [35], which expresses the degree of irregularity of the topology; and finally the 2-order Rényi entropy of a stationary Markovian random walk (H) [19].

In what follows, we describe the heat kernel and the derived invariants used to characterize the considered ensembles. The following material is principally reorganized from Ref. [38]. Let $G = (\mathcal{V}, \mathcal{E})$ be a graph with $n = |\mathcal{V}|$ vertices and $m = |\mathcal{E}|$ edges. Let $\mathbf{A}^{n \times n}$ be the adjacency matrix defined as $A_{ij} = 1$ if there is an edge between vertices $v_i, v_j \in \mathcal{V}$; $A_{ij} = 0$ otherwise. Let us define the degree of a vertex v_i as $\deg(v_i) = \sum_{j=1}^n A_{ij}$. In addition, let us define \mathbf{D} as a diagonal matrix of degree: $D_{ii} = \deg(v_i)$. Let \mathbf{L} be the Laplacian matrix given by $\mathbf{L} = \mathbf{D} - \mathbf{A}$; the normalized Laplacian matrix as $\hat{\mathbf{L}} = \mathbf{D}^{-1/2} \mathbf{L} \mathbf{D}^{-1/2}$. As a consequence, $\hat{\mathbf{L}}$ is symmetric and positive semi-definite and therefore it has non-negative eigenvalues. The eigendecomposition of the Laplacian is expressed as $\hat{\mathbf{L}} = \Phi \Lambda \Phi^T$, where Λ is the diagonal matrix containing the eigenvalues arranged as $0 = \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n \leq 2$; Φ contains the corresponding (unitary) eigenvectors as columns.

The heat equation [65] associated to $\hat{\mathbf{L}}$ is given by

$$\frac{\partial \mathbf{H}_t}{\partial t} = -\hat{\mathbf{L}} \mathbf{H}_t, \quad (1)$$

where \mathbf{H}_t is a doubly-stochastic $n \times n$ matrix, called heat matrix, and t is the time variable. Eq. 1 describes the diffusion of heat/information across the graph over time. Being doubly-stochastic, the heat matrix possesses a uniform stationary distribution. It is well-known that the solution to (1) is

$$\mathbf{H}_t = \exp(-\hat{\mathbf{L}}t), \quad (2)$$

which can be solved by exponentiating the spectrum of $\hat{\mathbf{L}}$:

$$\mathbf{H}_t = \Phi \exp(-\Lambda t) \Phi^T = \sum_{i=1}^n \exp(-\lambda_i t) \phi_i \phi_i^T. \quad (3)$$

The heat trace (HT) of \mathbf{H}_t is an invariant feature that is given by

$$\text{HT}(t) = \text{Tr}(\mathbf{H}_t) = \sum_{i=1}^n \exp(-\lambda_i t), \quad (4)$$

which thus takes into account only the eigenvalues of $\hat{\mathbf{L}}$. The heat content (HC) of \mathbf{H}_t is defined by considering also the eigenvectors of $\hat{\mathbf{L}}$:

$$\text{HC}(t) = \sum_{u \in \mathcal{V}} \sum_{v \in \mathcal{V}} \mathbf{H}_t(u, v) = \sum_{u \in \mathcal{V}} \sum_{v \in \mathcal{V}} \sum_{i=1}^n \exp(-\lambda_i t) \phi_i(v) \phi_i(u), \quad (5)$$

where with $\phi_i(v)$ we indicate the value related to the vertex v in the i th eigenvector.

Eq. 5 can be described in terms of power series expansion,

$$\text{HC}(t) = \sum_{m=0}^{\infty} q_m t^m. \quad (6)$$

By using the McLaurin series for the exponential function, we have

$$\exp(-\lambda_i t) = \sum_{m=0}^{\infty} \frac{(-\lambda_i)^m t^m}{m!}, \quad (7)$$

which substituted in Eq. 5 gives:

$$\text{HC}(t) = \sum_{u \in \mathcal{V}} \sum_{v \in \mathcal{V}} \sum_{i=1}^n \exp(-\lambda_i t) \phi_i(v) \phi_i(u) = \sum_{m=0}^{\infty} \sum_{u \in \mathcal{V}} \sum_{v \in \mathcal{V}} \sum_{i=1}^n \phi_i(v) \phi_i(u) \frac{(-\lambda_i)^m t^m}{m!}. \quad (8)$$

The q_m coefficients in (6) are graph invariants (called heat content invariants, HCI) that can be calculated in closed-form by using Eqs. 6 and 8:

$$q_m = \sum_{i=1}^n \left(\sum_{u \in \mathcal{V}} \phi_i(u) \right)^2 \frac{(-\lambda_i)^m}{m!}. \quad (9)$$

In what follows, we provide an argument to characterize the heat trace (4) as a property of a homogeneous ensemble of graphs. The heat trace (4) of a graph $G = (\mathcal{V}, \mathcal{E})$, with $n = |\mathcal{V}|$, can be expressed as

$$\text{HT}_G(t) = \sum_{i=1}^n \exp(-\lambda_i t) = 1 + \sum_{i=2}^n \exp(-\lambda_i t). \quad (10)$$

where λ_i are the eigenvalues of the normalized Laplacian of G . Let us define an ensemble \mathcal{C} of graphs, in which all graphs share a common characteristic spectral density. Such spectra can be synthetically described by considering the spectral density of \mathcal{C} . Accordingly, we can consider the eigenvalues as i.i.d. random variables, $\tilde{\lambda}_i$, assuming values according to the spectral density of the ensemble; note that $\tilde{\lambda}_1$ assumes deterministically the value 0. The HT of a generic graph $G \in \mathcal{C}$, $n = |\mathcal{V}|$, can be written as:

$$\text{HT}_G(t; n) = 1 + \sum_{i=2}^n \exp(-\tilde{\lambda}_i t) = 1 + \sum_{i=2}^n \exp(-\tilde{\lambda} t). \quad (11)$$

The last step in Eq. 11 is carried out by considering that, since the $\tilde{\lambda}_i$ are assumed as i.i.d., their values can be expressed as n realizations of a single random variable, $\tilde{\lambda}$, characterized by the same probability density function. For a fixed value of time t , we can define the ensemble HT, $\text{HT}_{\mathcal{C}}(n; t)$, as the mean HT over all graphs of the ensemble \mathcal{C} with varying size n . This quantity is given by:

$$\text{HT}_{\mathcal{C}}(n; t) = \langle \text{HT}_G(t; n) \rangle_{\mathcal{C}} = 1 + \sum_{i=2}^n \langle \exp(-\tilde{\lambda} t) \rangle_{\mathcal{C}} = 1 + (n-1) \langle \exp(-\tilde{\lambda} t) \rangle_{\mathcal{C}}. \quad (12)$$

Hence, $\text{HT}_{\mathcal{C}}(n; t)$ can be expressed as a linear function of the graph size

$$\text{HT}_{\mathcal{C}}(n; t) = 1 - \alpha_{\mathcal{C}}(t) + \alpha_{\mathcal{C}}(t) \cdot n \simeq \alpha_{\mathcal{C}}(t) \cdot n, \quad (13)$$

where $\alpha_{\mathcal{C}}(t) = \langle \exp(-\tilde{\lambda} t) \rangle_{\mathcal{C}} \in [0, 1]$ is a time-dependent slope that is characteristic for the entire ensemble \mathcal{C} .

References

- [1] J. A. Almendral and A. Díaz-Guilera. Dynamical and spectral properties of complex networks. *New Journal of Physics*, 9(6):187, 2007.
- [2] S. Balakrishnan, H. Kamisetty, J. G. Carbonell, S.-I. Lee, and C. J. Langmead. Learning generative models for protein fold families. *Proteins: Structure, Function, and Bioinformatics*, 79(4):1061–1078, 2011.

- [3] C. Baldassi, M. Zamparo, C. Feinauer, A. Procaccini, R. Zecchina, M. Weigt, and A. Pagnani. Fast and accurate multi-variate gaussian modeling of protein families: Predicting residue contacts and protein-interaction partners. *PloS one*, 9(3):e92721, 2014.
- [4] J. R. Banavar and A. Maritan. Physics of proteins. *Annual Review of Biophysics and Biomolecular Structure*, 36:261–280, 2007.
- [5] L. Bartoli, P. Fariselli, and R. Casadio. The effect of backbone on the small-world properties of protein contact maps. *Physical Biology*, 4(4):L1, 2007.
- [6] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.
- [7] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, Feb 2006. ISSN 03701573. doi: 10.1016/j.physrep.2005.10.009.
- [8] W. Boomsma, K. V. Mardia, C. C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamelryck. A generative, probabilistic model of local protein structure. *Proceedings of the National Academy of Sciences*, 105(26):8932–8937, 2008.
- [9] E. T. Bullmore and O. Sporns. The economy of brain network organization. *Nature Reviews Neuroscience*, 13(5):336–349, 2012. doi: 10.1038/nrn3214.
- [10] F. Chiti, N. Taddei, P. M. White, M. Bucciantini, F. Magherini, M. Stefani, and C. M. Dobson. Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nature Structural & Molecular Biology*, 6(11):1005–1009, 1999. doi: 10.1038/14890.
- [11] F. Chung. The heat kernel as the pagerank of a graph. *Proceedings of the National Academy of Sciences*, 104(50):19735–19740, 2007.
- [12] F. Comellas and J. Diaz-Lopez. Spectral reconstruction of complex networks. *Physica A: Statistical Mechanics and its Applications*, 387(25):6436–6442, Nov 2008. ISSN 03784371. doi: 10.1016/j.physa.2008.07.032.
- [13] L. d. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.
- [14] P. Csermely, K. Singh Sandhu, E. Hazai, Z. Hoksza, H. J M Kiss, F. Miozzo, D. V. Veres, F. Piazza, and R. Nussinov. Disordered proteins and network disorder in network descriptions of protein structure, dynamics and function: hypotheses and a comprehensive review. *Current Protein and Peptide Science*, 13(1):19–33, 2012.
- [15] A. Currin, N. Swainston, P. J. Day, and D. B. Kell. Synthetic biology for the directed evolution of protein biocatalysts: navigating sequence space intelligently. *Chemical Society Reviews*, pages –, 2015. doi: 10.1039/C4CS00351A.
- [16] A. Cuzzocrea, A. Papadimitriou, D. Katsaros, and Y. Manolopoulos. Edge betweenness centrality: A novel algorithm for qos-based topology control over wireless sensor networks. *Journal of Network and Computer Applications*, 35(4):1210–1217, 2012. doi: 10.1016/j.jnca.2011.06.001.
- [17] D. de Juan, F. Pazos, and A. Valencia. Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14(4):249–261, 2013.
- [18] S. C. de Lange, M. A. de Reus, and M. P. van den Heuvel. The laplacian spectrum of neural networks. *Frontiers in Computational Neuroscience*, 7, 2013. ISSN 1662-5188. doi: 10.3389/fncom.2013.00189.
- [19] M. Dehmer and A. Mowshowitz. A history of graph entropy measures. *Information Sciences*, 181(1):57–78, 2011. ISSN 0020-0255. doi: 10.1016/j.ins.2010.08.041.
- [20] L. Di Paola and A. Giuliani. Protein contact networks topology: a natural language for allostery. *Current Opinion in Structural Biology*, 2015.
- [21] L. Di Paola, M. De Ruvo, P. Paci, D. Santoni, and A. Giuliani. Protein contact networks: an emerging paradigm in chemistry. *Chemical Reviews*, 113(3):1598–1613, 2012.
- [22] M. Ekeberg, C. Lövkvist, Y. Lan, M. Weigt, and E. Aurell. Improved contact prediction in proteins: using pseudolikelihoods to infer potts models. *Physical Review E*, 87(1):012707, 2013.
- [23] E. Estrada. Universality in protein residue networks. *Biophysical Journal*, 98(5):890–900, 2010.
- [24] A. V. Finkelstein and O. V. Galzitskaya. Physics of protein folding. *Physics of Life Reviews*, 1(1):23 – 56, 2004. ISSN 1571-0645. doi: 10.1016/j.plrev.2004.03.001.

- [25] L. K. Gallos, H. A. Makse, and M. Sigman. A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks. *Proceedings of the National Academy of Sciences*, 109(8):2825–2830, 2012. doi: 10.1073/pnas.1106612109.
- [26] I. Gutman and B. Zhou. Laplacian energy of a graph. *Linear Algebra and its Applications*, 414(1):29–37, 2006.
- [27] T. A. Hopf, L. J. Colwell, R. Sheridan, B. Rost, C. Sander, and D. S. Marks. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell*, 149(7):1607–1621, 2012.
- [28] M. Ipsen and A. Mikhailov. Evolutionary reconstruction of networks. *Physical Review E*, 66(4):046109, Oct 2002. ISSN 1063-651X. doi: 10.1103/PhysRevE.66.046109.
- [29] B. Jana, F. Morcos, and J. N. Onuchic. From structure to function: the convergence of structure based models and co-evolutionary information. *Physical Chemistry Chemical Physics*, 16:6496–6507, 2014. doi: 10.1039/C3CP55275F.
- [30] H. Kamisetty, S. Ovchinnikov, and D. Baker. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence-and structure-rich era. *Proceedings of the National Academy of Sciences*, 110(39):15674–15679, 2013.
- [31] K. Kloster and D. F. Gleich. Heat kernel based community detection. *arXiv preprint arXiv:1403.3148*, 2014.
- [32] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- [33] R. Kühn and J. Van Mourik. Spectra of modular and small-world matrices. *Journal of Physics A: Mathematical and Theoretical*, 44(16):165205, 2011. doi: 10.1088/1751-8113/44/16/165205.
- [34] D. M. Leitner. Energy Flow in Proteins. *Annual Review of Physical Chemistry*, 59(1):233–259, 2008. doi: 10.1146/annurev.physchem.59.032607.093606. PMID: 18393676.
- [35] L. Livi and A. Rizzi. Graph ambiguity. *Fuzzy Sets and Systems*, 221:24–47, 2013. ISSN 0165-0114. doi: 10.1016/j.fss.2013.01.001.
- [36] L. Livi, A. Giuliani, and A. Rizzi. Toward a multilevel representation of protein molecules: comparative approaches to the aggregation/folding propensity problem. *ArXiv preprint arXiv:1407.7559*, Jul 2014.
- [37] L. Livi, A. Giuliani, and A. Sadeghian. Characterization of graphs for protein structure modeling and recognition of solubility. *arXiv preprint arXiv:1407.8033*, Jul 2014.
- [38] L. Livi, E. Maiorino, A. Pinna, A. Sadeghian, A. Rizzi, and A. Giuliani. Analysis of heat kernel highlights the strongly modular and heat-preserving structure of proteins. *ArXiv preprint arXiv:1409.1819*, Sep 2014.
- [39] E. Maiorino, L. Livi, A. Giuliani, A. Sadeghian, and A. Rizzi. Multifractal characterization of protein contact networks. *Physica A: Statistical Mechanics and its Applications*, 428:302–313, 2015. ISSN 0378-4371. doi: 10.1016/j.physa.2015.02.026.
- [40] D. S. Marks, L. J. Colwell, R. Sheridan, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander. Protein 3d structure computed from evolutionary sequence variation. *PloS one*, 6(12):e28766, 2011.
- [41] P. N. McGraw and M. Menzinger. Laplacian spectra as a diagnostic tool for network structure and dynamics. *Physical Review E*, 77:031102, Mar 2008. doi: 10.1103/PhysRevE.77.031102.
- [42] R. Merris. Laplacian matrices of graphs: a survey. *Linear Algebra and its Applications*, 10010:143–176, 1994.
- [43] R. Mishra and S. Bhushan. Knot theory in understanding proteins. *Journal of Mathematical Biology*, 65(6-7):1187–1213, 2012.
- [44] M. Mitrović and B. Tadić. Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities. *Physical Review E*, 80(2):026123, 2009.
- [45] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011.
- [46] F. Morcos, B. Jana, T. Hwa, and J. N. Onuchic. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proceedings of the National Academy of Sciences*, 110(51):20533–20538, 2013.
- [47] M. E. J. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.
- [48] M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

- [49] A. Onofrio, G. Parisi, G. Punzi, S. Todisco, M. A. Di Noia, F. Bossis, A. Turi, A. De Grassi, and C. L. Pierri. Distance-dependent hydrophobic-hydrophobic contacts in protein folding simulations. *Physical Chemistry Chemical Physics*, 16:18907–18917, 2014. doi: 10.1039/C4CP01131G.
- [50] M. Orozco. A theoretical view of protein dynamics. *Chemical Society Reviews*, 43:5051–5066, 2014. doi: 10.1039/C3CS60474H.
- [51] P. Paci, L. Di Paola, D. Santoni, M. De Ruvo, and A. Giuliani. Structural and functional analysis of hemoglobin and serum albumin through protein long-range interaction networks. *Current Proteomics*, 9(3):160–166, 2012. doi: 10.2174/157016412803251815.
- [52] T. P. Peixoto. Eigenvalue spectra of modular networks. *Physical Review Letters*, 111(9):098701, Aug 2013. ISSN 0031-9007. doi: 10.1103/PhysRevLett.111.098701.
- [53] R. C. Penner, M. Knudsen, C. Wiuf, and J. E. Andersen. Fatgraph models of proteins. *Communications on Pure and Applied Mathematics*, 63(10):1249–1297, 2010.
- [54] R. C. Penner, M. Knudsen, C. Wiuf, and J. E. Andersen. An algebro-topological description of protein domain structure. *PloS one*, 6(5):e19670, 2011.
- [55] M. Randić, J. Zupan, A. T. Balaban, D. Vikić-Topić, and D. Plavšić. Graphical representation of proteins. *Chemical Reviews*, 111(2):790–862, 2011.
- [56] H. D. Rozenfeld, C. Song, and H. A. Makse. Small-world to fractal transition in complex networks: a renormalization group approach. *Physical Review Letters*, 104:025701, Jan 2010. doi: 10.1103/PhysRevLett.104.025701.
- [57] P. Sah, L. O. Singh, A. Clauset, and S. Bansal. Exploring community structure in biological networks with random graphs. *BMC Bioinformatics*, 15(1):220, 2014. doi: 10.1186/1471-2105-15-220.
- [58] M. R. Segal. A novel topology for representing protein folds. *Protein Science*, 18(4):686–693, 2009.
- [59] M. J. Skwark, D. Raimondi, M. Michel, and A. Elofsson. Improved contact predictions using the recognition of protein like contact patterns. *PLoS Computational Biology*, 10(11):e1003889, 2014.
- [60] S. Tasdighian, L. Di Paola, M. De Ruvo, P. Paci, D. Santoni, P. Palumbo, G. Mei, A. Di Venere, and A. Giuliani. Modules identification in protein structures: the topological and geometrical solutions. *Journal of Chemical Information and Modeling*, 54(1):159–168, 2013. doi: 10.1021/ci400218v.
- [61] E. R. van Dam and W. H. Haemers. Developments on spectral characterizations of graphs. *Discrete Mathematics*, 309(3):576–586, 2009.
- [62] M. Vassura, L. Margara, P. Di Lena, F. Medri, P. Fariselli, and R. Casadio. Reconstruction of 3D structures from protein contact maps. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 5(3):357–367, 2008.
- [63] M. Vendruscolo, E. Paci, C. M. Dobson, and M. Karplus. Three key residues form a critical contact network in a protein folding transition state. *Nature*, 409(6820):641–645, 2001. doi: 10.1038/35054591.
- [64] P. G. W. Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie*, 30:50–56, 2014. ISSN 0300-9084. doi: 10.1016/j.biochi.2014.12.007.
- [65] B. Xiao, E. R. Hancock, and R. C. Wilson. Graph Characteristics from the Heat Kernel Trace. *Pattern Recognition*, 42(11):2589–2606, Nov. 2009. ISSN 0031-3203. doi: 10.1016/j.patcog.2008.12.029.
- [66] W. Yan, J. Zhou, M. Sun, J. Chen, G. Hu, and B. Shen. The construction of an amino acid network for understanding protein structure and function. *Amino Acids*, 46(6):1419–1439, 2014.
- [67] X. Zhang, R. R. Nadakuditi, and M. E. J. Newman. Spectra of random graphs with community structure and arbitrary degrees. *Physical Review E*, 89(4):042816, 2014.